BENCHMARKING TEST-TIME ADAPTATION UNDER ATTACK IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating test-time adaptation under attack requires specifying the targeted predictor, the images on which errors occur, and the cost of adaptation. We compare six adaptation methods and zero-shot classification across three vision–language backbones and fifteen datasets, using common weights, starting prompts, and input images within each task. The 1,890 evaluations cover clean inputs and five cached attacks. Their per-image predictions reveal distinctions that aggregate accuracy alone misses. MAC has higher mean accuracy than R-TPT over four attacks (52.96% versus 50.86%), but a lower fraction correct under every attack (31.63% versus 37.41%); this ordering reverses within 15 of the 45 tasks. R-TPT also takes 6.33 times as long on matched attacked inputs in an 18-task local comparison. DBD leads the cached suite, yet its above-clean gains depend strongly on the attack. A separate 100-image DTD study queries all seven predictors directly. DBD's 75.56% cached accuracy falls to 34.38% under adaptive search, but a control choosing between clean and cached inputs already yields 38.94%. Thus most of that gap removes a cached-input benefit; the remaining drop under the full attack is smaller. A matched ten-image check at five times the query budget measures further sensitivity, with substantial uncertainty. These results link method selection to attack coverage, error overlap, clean-correct retention, and measured cost, while keeping shared-input performance distinct from the bounded adaptive case study.

1 INTRODUCTION

Vision–language models classify images by matching visual representations to class descriptions (Radford et al., 2021; Zhai et al., 2023). Test-time adaptation (TTA) modifies prediction through prompt learning, aggregation of transformed views, or image-space counterattacks. Comparing these mechanisms under attack involves two decisions: which predictor the attacker targets, and which errors the evaluation counts. A perturbation optimized for the frozen backbone can have different effects after adaptation. Even on identical attacked images, higher average accuracy can coexist with failure on more images under at least one attack.

An average over attacks measures success under that mixture; requiring correctness under every selected attack measures its intersection. Neither describes preservation of correct clean predictions or adaptation cost. Differences in prompts, checkpoints, and attacked images further complicate method selection. We compare methods with common initial information, retaining their adaptation mechanisms and recording their compute budgets.

We evaluate TPT (Shu et al., 2022), R-TPT (Sheng et al., 2025), DBD (Liu et al., 2026), MTA (Zanella & Ben Ayed, 2024), TTC (Xing et al., 2025), MAC (Kim & Um, 2026), and zero-shot classification on three backbones and fifteen datasets. Each task shares pretrained weights, starting prompts, class order, and input bytes; adaptation resets between images. The 1,890 evaluations cover all 45 tasks at seed zero under clean inputs and five attacks. Four attacks share an ℓ_∞ budget; PGD-ℓ_2 is separate. Departures from native method configurations are specified in Section 3 and Appendix B.

Three findings illustrate the consequences. Square yields 56.71% accuracy for TPT, but PGD-10 yields 0.20%, making attack coverage decisive. MAC exceeds R-TPT in mean attacked accuracy, while R-TPT has higher per-image worst-case accuracy; their ordering reverses within 15 of 45

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tasks, and different attacks cause their exclusive failures. Matched hardware measurements add a substantial cost difference. Finally, a clean/cached selection control explains most of DBD's cached-to-adaptive gap on DTD, separating that effect from the remaining drop under the full attack.

Our contribution is a paired comparison of attack-specific errors, their overlap, clean retention, and measured cost across adaptation mechanisms. The balanced shared-input benchmark and separate seven-method adaptive case study connect changes in ranking to those observable differences. Composition checks identify where their magnitude depends on the chosen backbones and datasets.

2 Related work

Adaptation mechanisms. Test-time VLM methods modify prompts, features, or pixels. TPT minimizes entropy over selected augmented views for natural distribution shifts (Shu et al., 2022); R-TPT uses pointwise entropy and reliability-weighted ensembling for adversarial inputs (Sheng et al., 2025). MTA estimates a visual feature mode without prompt learning (Zanella & Ben Ayed, 2024), while DBD reconstructs features from transformation-induced directional bias (Liu et al., 2026). TTC optimizes input counterattacks, and MAC combines multiple views with corruption-aware weighting (Xing et al., 2025; Kim & Um, 2026). We compare these mechanisms through their predictions and processing costs under common inputs.

Comparative evaluation. TTAB identifies how model quality, model selection, and distribution shifts affect adaptation results (Zhao et al., 2023). R-TPT and DBD evaluate multiple datasets and attacks (Sheng et al., 2025; Liu et al., 2026). DBD also examines above-clean attacked accuracy, initially correct and incorrect samples, inference efficiency, and defense-aware attacks. Our benchmark studies how attack-specific errors and their overlap change method rankings across shared backbone–dataset tasks. Paired predictions connect these rankings to clean-correct retention, while matched hardware makes their cost differences measurable. The adaptive case study further separates gains on cached inputs from losses relative to clean inputs and from additional search, using a clean/cached initialization control.

Attack targets. AutoAttack and RobustBench standardize multi-attack evaluation (Croce & Hein, 2020; Croce et al., 2021). We use a specified cached suite with custom budgets and report its per-attack and joint correctness. Adaptive evaluation targets the defended computation (Tramèr et al., 2020). Our Square-style score queries (Andriushchenko et al., 2020) run each complete adaptation method, with separate randomness for search and final scoring. This tests a different predictor from the undefended backbone used to generate the common cached inputs.

Pretrained systems and configurations. CLIP, DataComp-trained OpenCLIP, and SigLIP differ in pretraining data and objectives (Radford et al., 2021; Gadre et al., 2023; Zhai et al., 2023). We compare their method configurations with shared initial prompts and information access, including explicit OpenCLIP/SigLIP extensions. Method-specific compute budgets remain part of the comparison. The shared-prompt DBD configuration omits CuPL. Reported rankings apply to these configurations, whose departures from native setups are specified in Section 3.

3 Benchmark design

Predictor and unit of comparison. For a fixed dataset prompt and class vocabulary, let f_b denote the zero-shot score function for backbone b, and $F_{b,m}(x;\xi)$ the class scores after method m adapts to image x with randomness ξ. Pretrained encoder weights stay fixed. TPT/R-TPT optimize temporary prompt context; MTA/DBD aggregate or reconstruct visual features; TTC/MAC optimize a temporary image correction. We compare these full predictors, resetting adaptation between images.

Shared inputs and method budgets. Within each clean or cached-attack task, methods share weights, class order, split, prompt, and image bytes. Each dataset uses one template per class for both attack generation and evaluation (Appendix A). TPT/R-TPT initialize four context tokens from that prompt and verify matching initial text features. A method receives only the current image and common model/prompt context: no labels, paired clean image, attack metadata, method-specific checkpoint, retrieval set, LLM description, or diffusion input. Adaptation resets between images; compute budgets differ (Table 1).

Table 1: Common-protocol settings. Views include the original image; budgets concern adaptation, not attack generation. Learning rates and thresholds are in Appendix B.

Method	Views	Adaptation budget	Adaptation rule
Zero-shot	1	No updates	Shared model and text features
TPT	64	1 prompt step	Entropy of mean class distribution
R-TPT	64	1 prompt step	Mean view entropy; reliability fusion
DBD	32	Feature reconstruction	Directional reconstruction; no CuPL
MTA	64	Iterative feature solver	Iterative feature aggregation; crop/flip
TTC	1	2 pixel steps, $4/255$	Raw-feature counterattack
MAC	2	4 pixel steps, $8/255$	Raw-feature multi-view counterattack

TTC/MAC's correction radii are adaptation budgets, distinct from the adversary's perturbation bound. DBD omits CuPL descriptions, and MTA omits optional CoOp and diffusion inputs. All methods use FP32 and fixed published/reference constants, without tuning on these test accuracies. Our OpenCLIP and SigLIP TPT/R-TPT extensions optimize prompts through their respective text encoders; SigLIP retains the CLIP-derived softmax-entropy objective. These shared-information configurations differ from the native runners, so the comparison does not claim to reproduce their original headline accuracies. Appendix B gives implementation details and official repositories.

Attack targets. With $\mathcal{A}_a$ denoting a specified attack algorithm and budget, we distinguish

$$x_i^a = \mathcal{A}_a(f_b, x_i, y_i), \qquad x_{i,m}^{\mathrm{ad}} = \mathcal{A}_{\mathrm{query}}(F_{b,m}, x_i, y_i). \tag{1}$$

Cached x_i^a is identical across methods: the attack uses the undefended model's gradients or scores, as appropriate, and never queries adaptation. Each method is evaluated on the stored image bytes, including their encoding precision. The adaptive search instead queries the full $F_{b,m}$ and produces a method-specific input; its protocol is in Section 6. Both attackers know y_i; neither labels nor paired clean images enter prediction on an attacked input.

The cached ℓ_∞ suite $\mathcal{A}$ comprises PGD-10, APGD-CE-30, APGD-T-30, and Square (5,000-query budget), all at $4/255$. APGD-T's iteration count applies to each targeted optimization. PGD-ℓ_2 (nominal optimization radius 1.0, 50 steps) is reported separately. PNG rounding can increase its decoded norm beyond 1.0 (Appendix C); its accuracy describes the quantized cached inputs. These custom budgets do not constitute standard AutoAttack; the two attack targets also answer different questions about transfer and adaptive robustness.

Models and coverage. We evaluate CLIP ViT-B/16, DataComp-trained OpenCLIP ViT-B/16, and SigLIP-B/16 on fifteen full test manifests (232,085 images; Appendix A). The primary grid contains $45 \times 7 \times 6 = 1{,}890$ backbone–dataset–method–condition combinations; all are reported at seed 0. Each of the 45 tasks contains all seven methods and six conditions, yielding 315 complete method families. Per-backbone and individual results appear in Appendix D; Appendix E adds CLIP-B/32 and CLIP-L/14.

Metrics. Within a backbone–dataset–method family, let c_i^a indicate correct classification of image i under attack a, and c_i^0 the same method's clean correctness. From the saved predictions,

$$\mathrm{RA}_a = \frac{100}{N}\sum_i c_i^a, \qquad \mathrm{MeanRA} = \frac{1}{|\mathcal{A}|}\sum_a \mathrm{RA}_a, \qquad \mathrm{MinRA} = \min_a \mathrm{RA}_a, \tag{2}$$

$$\mathrm{WCRA} = \frac{100}{N}\sum_i \prod_{a\in\mathcal{A}} c_i^a, \quad \mathrm{WCRA{+}C} = \frac{100}{N}\sum_i c_i^0 \prod_{a\in\mathcal{A}} c_i^a. \tag{3}$$

WCRA requires correctness under every selected attack; WCRA+C also requires clean correctness and retains all N images in the denominator. Hence $\mathrm{WCRA{+}C} \le \min(\mathrm{Clean}, \mathrm{WCRA})$ and $\mathrm{WCRA} \le \mathrm{MinRA}$. We compute metrics within each family, then weight each task by $1/45$ for every method. Every dataset contributes three backbones, so this also gives equal weight to each dataset. Test-set size does not affect task weight. Stochastic methods are evaluated through their recorded seed-0 predictions, not an exact expectation over randomness.

MeanRA gives accuracy under a uniform mixture of the selected attacks; WCRA gives the intersection of their successes. Neither describes clean utility, a different attack mixture, or full-method

Table 2: Shared 45-task accuracy comparison (%). MeanRA, WCRA and WCRA+C use the same four cached ℓ_∞ attacks. Retention is the fraction of each method's clean-correct images surviving all four, computed within families before averaging. Higher is better. PGD-ℓ_2 is separate, with nominal pre-encoding radius 1.0; PNG rounding can exceed that norm. 0.00 may be positive before rounding.

Method	Clean	MeanRA	PGD-ℓ_2	WCRA	WCRA+C	Retention
Zero-shot	67.81	0.98	0.40	0.00	0.00	0.00
TPT	69.09	15.92	0.53	0.00	0.00	0.00
R-TPT	68.51	50.86	55.73	37.41	37.20	50.82
DBD	67.55	**72.44**	**95.60**	**59.99**	**57.84**	**82.52**
MTA	**69.67**	42.94	43.51	28.06	28.04	36.54
TTC	58.92	20.77	5.92	0.49	0.48	0.73
MAC	66.90	52.96	66.81	31.63	31.52	43.72

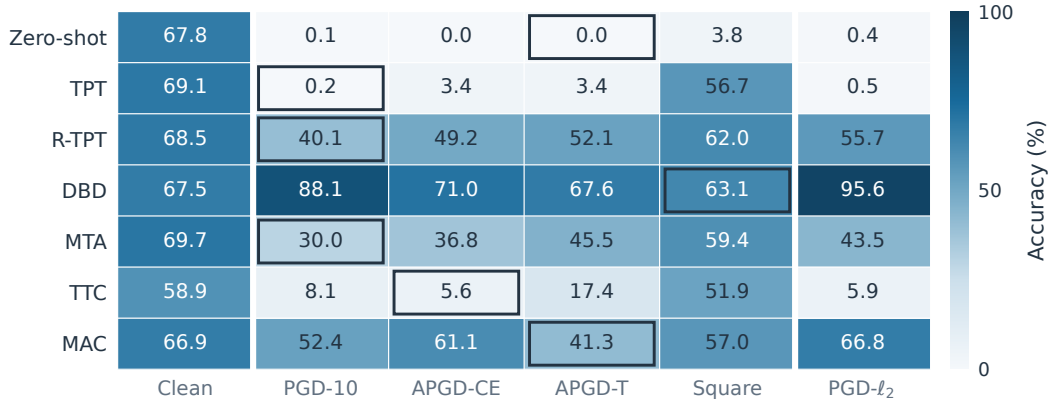

Figure 1: Accuracy by input condition over the same 45 tasks. Outlines mark each method's lowest aggregate accuracy among the four ℓ_∞ attacks. Clean and PGD-ℓ_2 are separate.

adaptive performance. In complete table columns, bold/shading marks the best value and underline the second distinct value; ties at printed precision share rank. These marks indicate ordering, not statistical significance.

Efficiency. We measure the 18 local tasks on one RTX 5090 with batch size one, FP32, and four Torch CPU threads. Timing includes preprocessing, transfers, adaptation, prediction, and CPU score copying, but excludes two warmup batches, model setup, and attack generation. We report clean and attacked steady-state means separately, with peak allocated GPU memory. Heterogeneous server GPU timings are excluded.

4 Results: Attack Identity and Aggregation

Table 2 compares the same 45 backbone–dataset tasks for every method; Figure 1 separates individual attacks.

Adaptation changes cached-attack transfer. Square yields 3.77% zero-shot accuracy versus 0.13% for PGD-10. TPT obtains 56.71% under Square but 0.20% under PGD-10, whereas DBD reaches 88.12% under PGD-10 and 63.06% under Square. These differences describe each predictor's response to the same cached inputs.

Across tasks, PGD-10 is a lowest-accuracy condition for R-TPT in 45 of 45 tasks; Square is lowest for DBD in 45; APGD-T is lowest for MAC in 42. An attack selected against the undefended source cannot represent every adapted predictor. Square alone favors TPT in a way its gradient-based

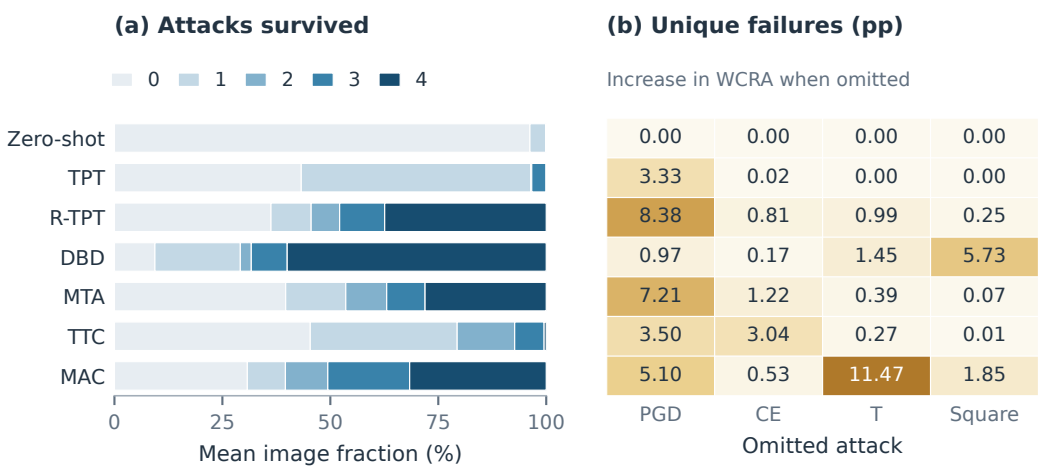

Figure 2: Per-image error structure, equally averaged over tasks. Left: fraction surviving exactly 0–4 attacks; the dark final segment equals WCRA. Right: WCRA increase when one attack is omitted, in percentage points. This is the fraction failing only that attack while surviving the other three. The four contributions sum to the "survives three" segment; they do not allocate failures shared by multiple attacks.

cached conditions do not support. This observation concerns the present adversarial comparison, not TPT's original natural-generalization problem.

Averages hide which images fail. MAC's MeanRA is 52.96%, above R-TPT's 50.86%, but its WCRA is 31.63%, below R-TPT's 37.41%. MAC's 41.17% MinRA leaves another 9.54 percentage points of failure exposed by intersecting predictions by image; R-TPT's MinRA-to-WCRA gap is 2.72 points. MeanRA measures average success over the chosen attacks, while WCRA requires survival of every recorded attacked version.

Within a task, let $K_i = \sum_{a \in \mathcal{A}} c_i^a$ count how many of the four attacked versions of image i are classified correctly. MeanRA measures $100\,\mathbb{E}[K_i]/4$, whereas WCRA measures $100\,\Pr(K_i = 4)$ for the empirical sample. Figure 2 averages these fractions equally over the 45 tasks. The fraction surviving exactly three attacks is 18.95% for MAC and 10.43% for R-TPT. Those three successes contribute to MeanRA, but the image fails the intersection.

The order-independent exclusive failure contribution of attack a is

$$U_a = \frac{100}{N} \sum_i (1 - c_i^a) \prod_{b \in \mathcal{A} \setminus \{a\}} c_i^b = \mathrm{WCRA}(\mathcal{A} \setminus \{a\}) - \mathrm{WCRA}(\mathcal{A}). \tag{4}$$

Removing APGD-T raises MAC's WCRA by 11.47 points, whereas removing PGD-10 raises R-TPT's by 8.38 points. DBD's largest unique contribution comes from Square, at 5.73 points. A small U_a can reflect failures shared with other suite members rather than an intrinsically weak attack; no cumulative attack ordering is needed.

Task-level comparisons retain the exceptions. Figure 3 compares R-TPT and MAC on every included task: MAC wins 25 under MeanRA, whereas R-TPT wins 35 under WCRA. The within-task ordering reverses in 15 cases. Requiring clean correctness yields 35 WCRA+C wins for R-TPT.

We estimate conditional image-sampling uncertainty with 10,000 paired bootstrap resamples within each fixed dataset, sharing image indices across backbones, methods, attacks, and metrics. Each task has weight $1/45$, with three backbones for every dataset. R-TPT minus MAC WCRA is 5.78 points with a 95% percentile interval [5.59, 5.98]; MAC minus R-TPT MeanRA is 2.10 points [1.97, 2.23]. These intervals condition on the evaluated systems and datasets; they do not estimate seed variability or generalization to new datasets. Shared indices prevent counting one image evaluated by three models as three independent samples.

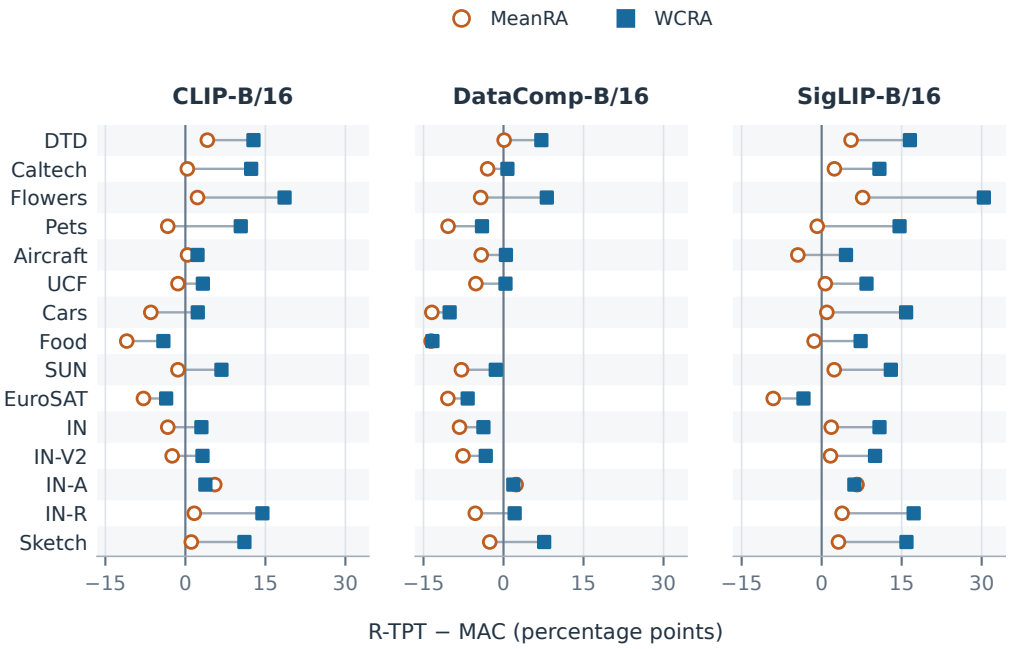

Figure 3: Paired R-TPT minus MAC differences on the 45 tasks. Positive values favor R-TPT. Hollow circles show MeanRA and filled squares show WCRA; connectors link two summaries of the same task. MAC wins 25 tasks in MeanRA; WCRA favors R-TPT on 35. These tasks are not independent random-seed replications.

Sensitivity to benchmark composition. Across fifteen dataset omissions and three backbone omissions, all 21 method-pair numerical orderings for MeanRA, WCRA, quantized PGD-ℓ_2, and retention persist. MTA remains the clean leader, despite five pairwise clean-ranking reversal events. Across backbone omissions, MAC's MeanRA advantage over R-TPT ranges from 0.03 to 3.84 points, while R-TPT's WCRA advantage ranges from 2.74 to 9.16. Without DataComp, the 0.03-point MeanRA difference is a near tie. These checks describe composition sensitivity, not uncertainty over new seeds or datasets (Appendix J).

5 Clean Behavior, Dataset Variation, and Inference Cost

Above-clean accuracy and clean retention differ. DBD leads the aggregate cached metrics, yet its MeanRA exceeds its own clean accuracy in 43 of 45 families. WCRA exceeds clean accuracy in 1. Thus average attacked accuracy and survival of the entire suite give different readings of the above-clean gain. Adding clean correctness changes DBD's aggregate from 59.99% WCRA to 57.84% WCRA+C.

Define $G_a = 100\,\mathbb{E}[(1-c_i^0)c_i^a]$ and $B_a = 100\,\mathbb{E}[c_i^0(1-c_i^a)]$, the fractions of all images changing from wrong to correct and from correct to wrong, respectively. They satisfy $\mathrm{RA}_a - \mathrm{Clean} = G_a - B_a$. Figure 4 reports these transitions relative to each method's own clean prediction, averaging the task-level fractions with the same equal weights. DBD gains 22.56 points and loses 1.99 under PGD-10, net 20.57 points. Under Square, it gains 2.61 points and loses 7.10, net -4.49 points. The opposing net changes arise from different transition patterns.

The attacks use the ground-truth label during generation; the defense sees only the resulting pixels. Above-clean accuracy may reflect structure in those pixels, as discussed in DBD's original study (Liu et al., 2026). These transitions identify gains in the cached comparison, without establishing a causal mechanism or showing that arbitrary perturbations improve recognition.

Conditional retention measures how much of the clean-correct set survives all four attacks: $R_C = 100\sum_i c_i^0 \prod_a c_i^a / \sum_i c_i^0$. We compute this ratio within each family before averaging. DBD retains

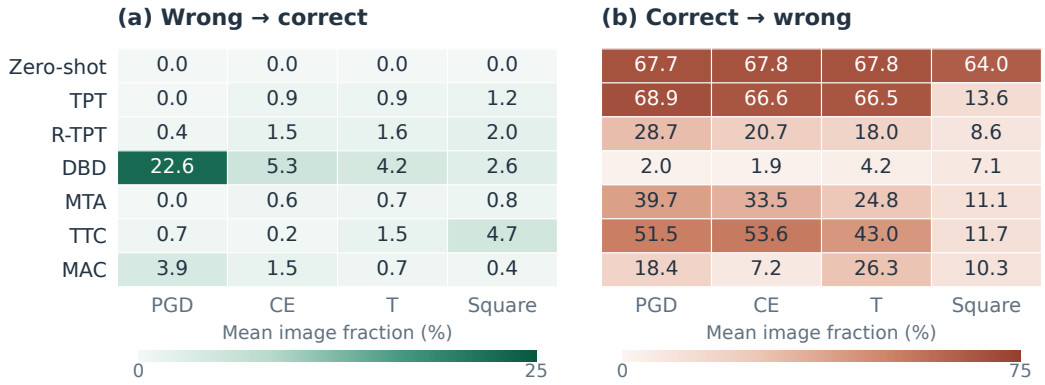

Figure 4: Prediction transitions relative to the same method on clean images. Entries are task-averaged percentages of images, with separate color scales for gains (0–25) and losses (0–75). Subtracting the right panel from the left recovers the attack-specific change in accuracy. These descriptive transitions do not establish recovery of naturally occurring errors.

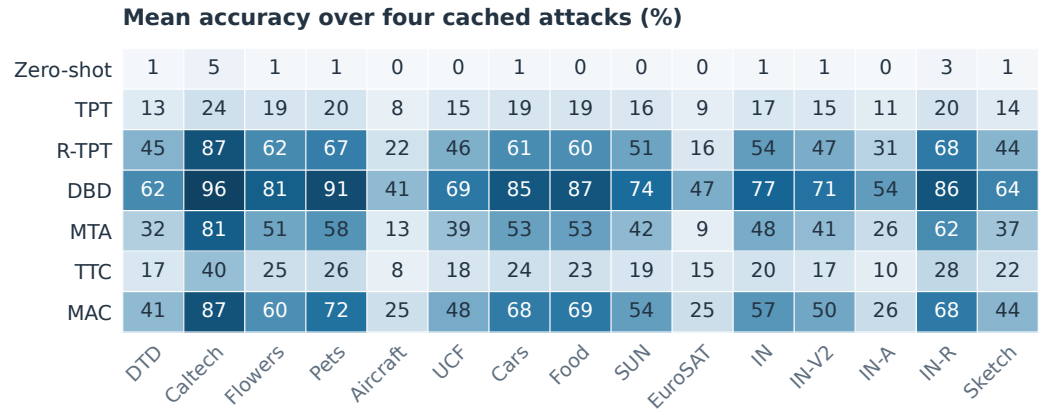

Figure 5: MeanRA over all fifteen datasets, rounded to whole percentages for legibility. Each column averages the same three backbones. Exact individual results are reported in the appendix.

82.52% of its clean-correct predictions, R-TPT 50.82%, MAC 43.72%, and MTA 36.54%. Each method has its own clean-correct subset, so retention compares different image sets across methods. WCRA+C instead uses all images as its denominator; both measures complement clean accuracy.

Performance across datasets. Appendix D reports all individual results and a Clean/PGD-10/APGD-CE panel on the same fifteen datasets and three backbones.

MTA has the highest clean accuracy, 69.67%, compared with 67.81% for zero-shot and 69.09% for TPT. Mean accuracy over PGD-10 and APGD-CE is instead highest for DBD at 79.56%, followed by MAC at 56.78% and R-TPT at 44.69%. Clean and cached-input leaders therefore differ even in this panel; two-attack means still differ from WCRA.

Figure 5 shows dataset variation averaged across included backbones. A separate comparison adds CLIP-B/32 and CLIP-L/14 on the six local datasets under the same two attacks (Appendix E). These comparisons across pretrained systems and domains do not isolate architecture.

Timing depends on the input condition. Table 3 and Figure 6 use the six-dataset RTX 5090 cohort. Both accuracy–cost axes cover these same 18 tasks; server timing and memory are excluded. On cached ℓ_∞ inputs, R-TPT requires 338.32 ms per image versus 53.45 ms for MAC, a 6.33-fold ratio. MAC also has higher local MeanRA (55.63% versus 54.88%), whereas R-TPT has higher WCRA

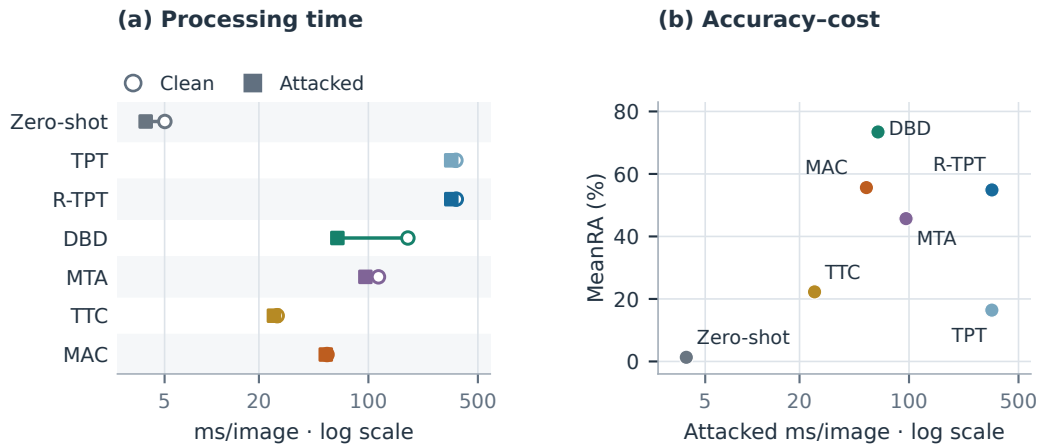

Figure 6: Matched local accuracy and processing cost on six datasets (18 tasks). Left: clean and four-attack mean latency, on a logarithmic scale. Right: MeanRA versus attacked latency on the same GPU. Connecting two input conditions does not assert that a particular internal operation causes their timing difference.

Table 3: Local mean steady-state processing time and mean run-level peak allocated GPU memory. Attacked values average the four ℓ_∞ conditions; setup and attack generation are excluded. Lower is better. Method budgets differ.

Method	Clean ms/img	ℓ_∞ ms/img	Mean peak MiB
Zero-shot	**5.02**	**3.79**	**681.41**
TPT	362.04	337.86	2182.61
R-TPT	364.04	338.32	2182.61
DBD	178.73	63.38	938.35
MTA	115.95	95.70	1239.43
TTC	26.18	24.94	798.06
MAC	54.45	53.45	913.38

(43.15% versus 34.39%). DBD's 63.38 ms attacked time accompanies the highest local cached-suite MeanRA and WCRA; its clean time is 178.73 ms. On Aircraft, DBD's clean/attacked times are 427.16/62.76 ms, whereas on Caltech they are 82.37/63.90 ms. A single pooled latency therefore cannot estimate deployment cost under a different input mix.

TPT/R-TPT require more peak allocated memory than MAC and DBD (Table 3). Allocated memory excludes total GPU occupancy; hardware is matched while compute budgets remain method-specific.

6 Adaptive attack against the full method

Attack protocol. We attack each full CLIP-B/16 predictor for all seven methods on the same 100 DTD images. A Square-style search minimizes the true-versus-best-other score margin within $4/255$, averaging four method draws per candidate. Every query resets adaptation and includes stochastic views and updates, without differentiating through the optimizer. The cached PGD-10 image is an initial candidate.

Scoring and budgets. Sixteen disjoint method seeds score clean, cached, and selected adaptive images; none selects attack candidates. We report accuracy and paired clean-correct loss with exact counts. The budget comparison uses the same ten IDs at both settings, separately from the 100-image mean. Query counts exclude cached warm-start generation. Appendix G gives full settings and the separate eight-image pilot. On the matched ten images, the 500-candidate configuration lowers DBD

Table 4: Full-method adaptive accuracy (%) on DTD. Left: 100 images, 100 candidates (400 method calls/image). Right: the same ten-image subset at 100/500 candidates (400/2,000 calls/image). Accuracy averages 16 held-out draws/image; CC-loss is the percentage of paired clean-correct draws made incorrect (lower is better). Bold/shading and underline mark the best and second distinct values within each column.

	100 images, 100 candidates				Matched 10 images	
Method	Clean	Cached	Adaptive	CC-loss	100 cand.	500 cand.
Zero-shot	42.00	0.00	0.00	100.00	0.00	0.00
TPT	42.69	0.00	0.00	100.00	0.00	0.00
R-TPT	**46.06**	31.13	29.44	37.04	**9.38**	**9.38**
DBD	40.00	**75.56**	**34.38**	**14.84**	8.75	2.50
MTA	41.88	14.12	14.25	65.97	0.00	0.00
TTC	42.31	4.94	3.62	91.43	0.00	0.00
MAC	41.12	26.88	23.56	44.98	6.25	3.75

accuracy from 8.75% to 2.50% and MAC from 6.25% to 3.75%; R-TPT stays at 9.38%. Zero-shot, TPT, MTA, and TTC remain at zero. DBD's decrease occurs on one image; both decreases have image-bootstrap intervals containing zero. The proposal schedule also changes, so this diagnostic neither isolates budget alone nor establishes convergence.

Adaptive accuracy and clean-correct failures. DBD changes from 75.56% on cached PGD-10 to 34.38% under adaptive search, compared with 40.00% clean. Of its 640 clean-correct held-out draws, 95 become incorrect. The cached-input gain does not persist under this search; clean-correct losses provide a separate measure of harm. A two-candidate control selects clean or cached PGD-10 using search margins (eight full-method calls), then uses held-out scoring. For DBD it selects clean on 92 images and obtains 38.94%; the further drop under full search is 4.56 points [1.44, 8.50], versus 5.62 [2.00, 10.06] relative to clean. MAC's control-relative drop is 2.19 [-0.25, 5.06]. These 95% intervals resample the 100 images with all draws kept together (Appendix H). The control excludes the random-stripe initial candidate. The contrast therefore measures the joint effect of stripe initialization and subsequent search; saved held-out predictions cannot separate those contributions.

7 Discussion and conclusions

Method selection depends on which inputs, failures, and costs matter. MTA leads clean accuracy and DBD the cached suite; MAC offers higher mean attacked accuracy and lower latency than R-TPT, while R-TPT has higher per-image worst-case accuracy. Per-image errors explain this reversal. On the adaptive case study, a clean/cached selection control removes most of DBD's cached-input gain; full search causes additional damage.

These findings support reporting per-attack accuracy, its suite mean, per-image intersection, and clean-correct retention together, with cost measured on matched hardware and inputs. Method-aware attacks additionally need a clean reference and controls for their starting candidates.

The results concern fixed datasets, backbones, and one method seed; image bootstraps measure within-task sampling uncertainty. Common prompts and backbone extensions change some native setups without isolating each change's effect; compute budgets remain unequal. The adaptive study covers one dataset/backbone, with a ten-image budget-and-schedule diagnostic and no convergence claim. Quantized PGD-ℓ_2 inputs can exceed the nominal radius. Within this scope, the benchmark connects method rankings to measured error patterns and processing cost.

REFERENCES

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*, pp. 484–501. Springer, 2020.

Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, pp. 446–461. Springer, 2014. URL https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3606–3613, 2014.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pp. 2206–2216. PMLR, 2020.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: A standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. URL https://openreview.net/forum?id=SSKZPJCt7B.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. URL https://www.image-net.org/static_files/papers/imagenet_cvpr09.pdf.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pp. 178–178. IEEE, 2004.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, et al. DataComp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2304.14108.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. URL https://arxiv.org/abs/1709.00029.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a. URL https://openaccess.thecvf.com/content/ICCV2021/html/Hendrycks_The_Many_Faces_of_Robustness_A_Critical_Analysis_of_Out-of-Distribution_ICCV_2021_paper.html.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b. URL https://openaccess.thecvf.com/content/CVPR2021/html/Hendrycks_Natural_Adversarial_Examples_CVPR_2021_paper.html.

Sunoh Kim and Daeho Um. When CLIP sees more, it fights back harder: Multi-view guided adaptive counterattacks for test-time adversarial robustness. *arXiv preprint arXiv:2606.06938*, 2026. URL https://arxiv.org/abs/2606.06938.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 554–561, 2013. URL https://cs.stanford.edu/groups/vision/pdf/3drr13.pdf.

Liangsheng Liu, Si Chen, Jiamin Wu, Weiwei Feng, Zhixin Cheng, Xiaotian Yin, Wenfei Yang, and Tianzhu Zhang. Adversarial attacks already tell the answer: Directional bias-guided test-time defense for vision-language models. *arXiv preprint arXiv:2606.06186*, 2026.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian conference on computer vision, graphics & image processing*, pp. 722–729. IEEE, 2008.

Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pp. 3498–3505. IEEE, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5389–5400. PMLR, 2019. URL https://proceedings.mlr.press/v97/recht19a.html.

Lijun Sheng, Jian Liang, Zilei Wang, and Ran He. R-tpt: Improving adversarial robustness of vision-language models through test-time prompt tuning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29958–29967, 2025.

Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 2020. URL https://arxiv.org/abs/2002.08347.

Haohan Wang, Songwei Ge, Eric P. Xing, and Zachary C. Lipton. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper/9237-learning-robust-global-representations-by-penalizing-local-predictive-power.pdf.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492, 2010. URL https://3dvision.princeton.edu/projects/2010/SUN/.

Songlong Xing, Zhengyu Zhao, and Nicu Sebe. Clip is strong enough to fight back: Test-time counterattacks towards zero-shot adversarial robustness of clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15172–15182, 2025.

Maxime Zanella and Ismail Ben Ayed. On the test-time zero-shot generalization of vision-language models: Do we really need prompt learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23783–23793, 2024.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Hao Zhao, Yuejiang Liu, Alexandre Alahi, and Tao Lin. On pitfalls of test-time adaptation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42058–42080. PMLR, 2023. URL https://proceedings.mlr.press/v202/zhao23d.html.

REPRODUCIBILITY STATEMENT

Dataset splits, starting prompts, adaptation hyperparameters, and attack budgets are specified in the experimental protocol and appendices. Metrics are computed from per-image class scores with aligned sample identities. The accompanying analysis includes the correctness masks, paired-bootstrap procedure, and scripts used to generate tables and figures.

AI USE STATEMENT

Generative AI tools were used to assist in drafting portions of the manuscript and to improve the clarity, grammar, and presentation of the writing. All AI-assisted text was reviewed, revised, and verified by the authors. The authors take full responsibility for the final content of the paper.

A Dataset manifests and exact starting prompts

All completed runs cover their full configured test manifests. The fifteen datasets contain 232,085 test images: 17,405 in the six local manifests and 214,680 in the nine server manifests. The table records observed vocabulary sizes, including 100 classes for Caltech101 and dense 200-class vocabularies for ImageNet-A/R. These are the actual evaluation manifests rather than assumed canonical training/test sizes. UCF101 uses still images, not a temporal video model. Class IDs, names, and order remain available in recorded results.

Table 5: All fifteen test manifests and shared initial prompt templates. The six local datasets use the recorded Zhou-style splits; the remaining rows retain their configured dataset-specific manifests.

Dataset	Images	Classes	Initial prompt
DTD (Cimpoi et al., 2014)	1692	47	a photo of a {class} texture
Caltech101 (Fei-Fei et al., 2004)	2465	100	a photo of a {class}
Flowers102 (Nilsback & Zisserman, 2008)	2463	102	a photo of a {class}, a type of flower
OxfordPets (Parkhi et al., 2012)	3669	37	a photo of a {class}, a type of pet
FGVCAircraft (Maji et al., 2013)	3333	100	a photo of a {class}, a type of aircraft
UCF101 (Soomro et al., 2012)	3783	101	a photo of a {class}
StanfordCars (Krause et al., 2013)	8041	196	a photo of a {class}
Food101 (Bossard et al., 2014)	30300	101	a photo of a {class}
SUN397 (Xiao et al., 2010)	19850	397	a photo of a {class}
EuroSAT (Helber et al., 2019)	8100	10	a remote sensing image of {class}
ImageNet (Deng et al., 2009)	50000	1000	a photo of a {class}
ImageNetV2 (Recht et al., 2019)	10000	1000	a photo of a {class}
ImageNetA (Hendrycks et al., 2021b)	7500	200	a photo of a {class}
ImageNetR (Hendrycks et al., 2021a)	30000	200	a photo of a {class}
ImageNetSketch (Wang et al., 2019)	50889	1000	a photo of a {class}

B Implementation details and departures from native runners

Detailed budgets are fixed throughout: TPT and R-TPT use 64 views, lowest-entropy 10% selection, four context tokens, one AdamW update at learning rate 0.005, and weight decay 0.01. R-TPT uses 20 fusion neighbors and temperature 0.01. DBD uses 32 total views, 16 selected transformed views, threshold 0.8, and direction multiplier 2.5. MTA uses $\lambda_y = 0.2$, $\lambda_q = 4$, five iterations per loop, tolerance 10^{-6}, and neighborhood fraction 0.3. TTC uses two counterattack steps of $1/255$ within $4/255$, $\beta = 2$, and threshold 0.2. MAC uses two views, four steps of $2/255$ within $8/255$, threshold 0.7, and temperature 0.01. TTC and MAC use random initialization.

The main experiment uses common pretrained weights and one dataset template per class. Model and preprocessing settings are shared within each backbone–dataset task. This choice removes method-specific prompt descriptions and checkpoints from the information budget. It changes the native setup of some methods and is part of the experimental question rather than an assertion of exact original-paper reproduction.

TPT (Shu et al., 2022) and R-TPT (Sheng et al., 2025) optimize four continuous prefix positions while preserving class tokens, suffixes, punctuation, masks, and model-specific pooling. Initial text features are checked against the common prompt bank. Image features are reused within each sample; context and optimizer reset between samples. Their distinct entropy objectives and R-TPT fusion are retained. Self-neighbors are removed by index, including duplicate-view cases.

DBD (Liu et al., 2026) retains its direction-based reconstruction and reference transformation recipe but omits CuPL descriptions. MTA (Zanella & Ben Ayed, 2024) uses its solver with crop/flip views;

optional CoOp prompts, diffusion images, and its runner's TPT/AugMix setting are not used. Degenerate solver inputs have numerical guards. TTC (Xing et al., 2025) and MAC (Kim & Um, 2026) compute counterattack updates from raw projected features and normalize for classification. They never access a paired clean image during attacked inference.

OpenCLIP uses its native text embeddings, causal masking, pooling, and projection. SigLIP preserves its tokenizer, padding, pooling, and logit scale/bias; its prompt optimization uses the CLIP-derived softmax-entropy objective. These are explicit extensions. All local timing is FP32 rather than a reproduction of each native runner's precision policy. A-TPT is outside the common grid.

Table 6: Official repositories for the evaluated adaptation methods.

Method	Official repository
TPT	`azshue/TPT`
R-TPT	`TomSheng21/R-TPT`
DBD	`liuls2002/DBD`
MTA	`MaxZanella/MTA`
TTC	`Sxing2/CLIP-Test-time-Counterattacks`
MAC	`sunoh-kim/MAC`

C Reproducibility and Implementation Checks

Pretrained models and image geometry. Table 7 identifies the public checkpoints. All use 224×224 RGB inputs. CLIP and DataComp use bicubic resizing of the shorter edge to 224 followed by a center crop. SigLIP resizes directly to 224×224. CLIP/DataComp normalization uses mean $(0.48145466, 0.4578275, 0.40821073)$ and standard deviation $(0.26862954, 0.26130258, 0.27577711)$; SigLIP uses 0.5 for both in each channel. The attack acts on the resulting spatially processed tensor in $[0, 1]$, with normalization inside the source classifier. Thus the perturbation domain is the model-resolution image, not the original-resolution JPEG.

Table 7: Checkpoint identifiers. The last two backbones appear only in the supplementary five-backbone comparison.

Backbone	Checkpoint
CLIP-B/16	`openai/clip-vit-base-patch16`
DataComp-B/16	OpenCLIP `ViT-B-16`, pretrained tag `datacomp_xl_s13b_b90k`
SigLIP-B/16	`google/siglip-base-patch16-224`
CLIP-B/32	`openai/clip-vit-base-patch32`
CLIP-L/14	`openai/clip-vit-large-patch14`

Cached attack parameters. Nominal optimization radii are $4/255$ for the four ℓ_∞ attacks and 1.0 for PGD-ℓ_2. The recorded configurations specify seed 0, one restart, and one EOT sample, with no attack-specific parameter overrides. Table 8 spells out the corresponding generator settings. APGD adapts its step size; its initial step is 2ϵ and oscillation threshold is 0.75. APGD-T uses the targeted difference-of-logits-ratio objective and tries class ranks 2–10 of the clean source prediction, stopping once the source is misclassified. Its 30-step budget is per target, with one restart per target; it does not search every class in a 1,000-class vocabulary. Square uses margin loss, an initial square fraction of 0.8, a schedule rescaled to its query budget, and random stripe initialization. Successful source misclassification can stop query search early.

The retained attack implementation resolves APGD-T to the dedicated targeted DLR routine; its generic configuration loss field is not used by that routine.[1] The historical caches did not store every resolved constructor argument or the installed attack-library version. We therefore distinguish the

[1]The APGD-T and Square implementations specify the corresponding attack-library defaults.

Table 8: Generator recipe for the five reported cached attacks. Step sizes are in unnormalized $[0, 1]$ image coordinates. The final column supplements the recorded budgets with constructor settings from the retained generator.

Attack	Budget	Objective and initialization
PGD-ℓ_∞	10 steps	Cross-entropy; step $2/255$; one random start
APGD-CE	30 steps	Cross-entropy; adaptive step; one random start
APGD-T	$30\times \leq 9$ steps	Targeted DLR; adaptive step; one start per target
Square	5,000 queries	Margin; $p_0 = 0.8$; one random start
PGD-ℓ_2	50 steps	Cross-entropy; step $1/12$; one random start

saved budgets from the code-derived defaults in Table 8; the latter are not a per-cache replay of the historical software environment. The expanded attack specification is available for 224 of the 225 primary caches. The exception, SigLIP/ImageNetSketch Square, retains the attack name, radius, and seed but not the expanded parameter record. All seven methods are evaluated on the same stored images; reconstructing this cache's complete generation settings remains limited by that missing record.

Image serialization. The generated tensor is clamped to $[0, 1]$, rounded to the nearest integer after multiplication by 255, and stored as an 8-bit lossless PNG. All methods decode the same cached bytes. Reported radii are the bounds used during continuous optimization; the cache's stored perturbation measurements precede PNG quantization. Rounding does not preserve the original ℓ_2 ball: a check of 16 evenly spaced images from the 1,692-image CLIP-B/16 DTD cache finds 14 with decoded perturbation norm above 1.0, with maximum 1.087. This diagnostic covers one cache and does not estimate the violation frequency over the suite. The PGD-ℓ_2 column therefore denotes a nominal optimization radius, not strict decoded-image robustness at radius 1.0. Its common-input method comparison remains well defined.

What the implementation checks establish. Independent numerical fixtures compare wrapper calculations with functions extracted from the official implementations: TPT/R-TPT entropy losses and gradients, R-TPT fusion, DBD's transformed views and directional reconstruction, MTA's solver, and TTC/MAC's raw-feature counterattack updates and aggregation. Additional checks exercise prompt initialization, context gradients, frozen backbone weights, common original views, and episodic resets on small CLIP, SigLIP, and OpenCLIP models. These CPU checks pass. Separate pretrained text checks on three classes per prompt give maximum absolute initialization errors of 1.20×10^{-7} for CLIP-B/16 and 5.97×10^{-8} for SigLIP-B/16. They support the implemented mechanisms and shared starting point; they do not establish full-dataset equality with native runners. We have not isolated the accuracy effects of replacing native prompt descriptions or optional checkpoints, and the OpenCLIP/SigLIP extensions should be read with that limitation.

D PER-BACKBONE AGGREGATES AND INDIVIDUAL ACCURACIES

The primary suite averages the same 45 backbone–dataset tasks for every method. All 315 method families contain clean predictions and the five attack conditions, giving 1,890 evaluations. The tables below report per-backbone aggregates and all individual accuracies on this complete grid.

Table 9: The complete 45-task comparison on clean, PGD-10, and APGD-CE (945 runs). Accuracies (%) are equally averaged over all fifteen datasets and three backbones. This table uses the same tasks as the primary comparison and isolates two of its four ℓ_∞ attacks. Higher is better; bold/shading and underline identify the best and second distinct values.

Method	Clean	PGD-10	APGD-CE
Zero-shot	67.81	0.13	0.01
TPT	<u>69.09</u>	0.20	3.36
R-TPT	68.51	40.14	49.24
DBD	67.55	**88.12**	**71.00**
MTA	**69.67**	30.03	36.83
TTC	58.92	8.14	5.57
MAC	66.90	<u>52.42</u>	<u>61.13</u>

Table 10: Per-backbone primary-suite aggregates (%). Every backbone and method covers the same fifteen datasets. Differences between pretrained systems do not isolate architecture or pretraining data. Best and second best are marked within each block/column; DBD uses shared prompts without CuPL.

Method	Clean	MeanRA	MinRA	WCRA	WCRA+C	PGD-ℓ_2
CLIP-B/16						
Zero-shot	62.77	0.23	0.00	0.00	0.00	0.06
TPT	<u>64.66</u>	16.00	0.08	0.00	0.00	0.09
R-TPT	64.31	43.46	<u>31.04</u>	<u>27.79</u>	<u>27.64</u>	49.45
DBD	63.41	**68.15**	**58.41**	**54.49**	**52.80**	**94.26**
MTA	**64.86**	32.35	18.34	16.41	16.41	32.23
TTC	56.90	18.35	3.81	0.24	0.24	3.61
MAC	62.56	<u>44.91</u>	30.80	21.36	21.33	<u>62.12</u>
DataComp-B/16						
Zero-shot	69.10	0.87	0.00	0.00	0.00	0.09
TPT	<u>70.52</u>	16.93	0.12	0.00	0.00	0.27
R-TPT	70.49	52.62	41.45	38.66	38.48	57.66
DBD	69.93	**74.61**	**65.08**	**62.53**	**60.69**	**96.70**
MTA	**71.27**	45.54	32.04	30.26	30.21	47.89
TTC	63.93	21.75	4.07	0.41	0.41	6.03
MAC	67.51	<u>58.85</u>	<u>48.89</u>	<u>39.62</u>	<u>39.36</u>	<u>75.71</u>
SigLIP-B/16						
Zero-shot	71.57	1.83	0.00	0.00	0.00	1.05
TPT	<u>72.08</u>	14.84	0.20	0.00	0.00	1.23
R-TPT	70.73	<u>56.51</u>	<u>47.91</u>	<u>45.79</u>	<u>45.48</u>	60.08
DBD	69.30	**74.57**	**65.67**	**62.95**	**60.02**	**95.84**
MTA	**72.87**	50.94	39.59	37.51	37.50	50.40
TTC	55.93	22.19	5.17	0.82	0.80	8.13
MAC	70.61	55.12	43.82	33.91	33.87	<u>62.60</u>

Table 11: All 45 backbone–dataset tasks. Accuracy (%) for clean, PGD-10, APGD-CE, APGD-T, Square and PGD-ℓ_2, plus the four-attack WCRA. PGD-ℓ_2 has nominal pre-encoding radius 1.0; decoded norms may exceed it (Appendix C). Bold/shading: best; underline: second best within each seven-method column and task. Ties share rank at displayed precision.

Dataset	Method	Clean	PGD	CE	T	Square	ℓ_2	WCRA
CLIP-B/16								
DTD	Zero-shot	44.62	0.00	0.00	0.00	0.95	0.18	0.00
	TPT	45.92	0.12	3.55	3.72	38.59	0.18	0.00
	R-TPT	**46.63**	28.13	33.81	34.69	**40.43**	39.72	25.41
	DBD	45.27	**76.65**	**47.10**	**43.79**	39.89	**88.65**	**37.71**
	MTA	44.33	13.71	12.53	18.79	33.81	17.55	10.87
	TTC	42.02	2.54	5.38	10.05	34.87	4.73	0.35
	MAC	43.97	28.61	37.88	19.39	34.69	45.09	12.65
Caltech101	Zero-shot	91.97	0.32	0.00	0.00	5.44	0.61	0.00
	TPT	**93.31**	0.45	2.11	2.47	87.71	0.77	0.00
	R-TPT	92.86	72.29	76.88	80.65	89.41	87.10	68.68
	DBD	93.23	**98.46**	**94.04**	**92.45**	**91.44**	**99.76**	**89.70**
	MTA	92.62	51.52	59.43	73.31	88.28	74.97	49.45
	TTC	92.21	16.11	6.25	29.29	85.88	20.04	1.10
	MAC	92.45	76.47	87.30	67.38	86.61	91.48	56.35
Flowers102	Zero-shot	70.89	0.00	0.00	0.00	0.49	0.00	0.00
	TPT	**72.15**	0.00	2.84	2.88	64.43	0.00	0.00
	R-TPT	71.17	36.26	48.07	50.10	67.44	60.09	33.29
	DBD	71.66	**89.57**	**71.70**	**69.79**	**68.17**	**96.91**	**62.93**
	MTA	70.89	17.74	21.11	35.36	62.16	36.91	15.35
	TTC	67.24	1.75	4.43	11.17	60.33	1.38	0.00
	MAC	71.50	42.14	63.78	22.29	64.56	62.36	14.70
OxfordPets	Zero-shot	87.63	0.00	0.00	0.00	0.57	0.00	0.00
	TPT	87.95	0.00	2.53	2.53	78.36	0.00	0.00
	R-TPT	88.12	41.73	55.27	60.15	82.45	70.67	38.02
	DBD	88.23	**96.13**	**87.68**	**88.20**	**84.33**	**97.19**	**80.29**
	MTA	**88.50**	27.23	38.76	50.07	79.61	54.59	24.53
	TTC	77.95	6.62	6.57	11.77	71.38	1.39	0.30
	MAC	88.23	56.96	77.02	39.63	79.20	75.36	27.64
FGVCAircraft	Zero-shot	24.75	0.00	0.00	0.00	0.00	0.00	0.00
	TPT	23.91	0.00	5.91	5.97	19.89	0.00	0.00
	R-TPT	24.60	10.53	17.28	17.70	21.84	16.08	7.74
	DBD	26.73	**58.36**	**26.91**	**24.06**	**23.34**	**66.67**	**18.36**
	MTA	**26.76**	2.52	5.19	7.23	17.61	4.92	1.86
	TTC	15.39	0.54	3.75	5.88	14.67	0.00	0.03
	MAC	24.18	18.45	20.31	9.39	17.55	28.92	5.46
UCF101	Zero-shot	63.84	0.00	0.00	0.00	0.66	0.00	0.00
	TPT	65.72	0.00	4.41	4.47	59.29	0.00	0.00
	R-TPT	65.74	35.47	45.86	48.16	62.20	52.21	32.43
	DBD	65.69	**87.13**	**68.28**	**65.45**	**63.05**	**96.14**	**60.61**
	MTA	**66.51**	20.54	29.02	38.22	59.45	38.12	18.69
	TTC	64.76	4.86	4.84	13.69	54.32	1.90	0.13
	MAC	64.71	46.39	55.33	39.39	56.15	65.42	29.16

Continued on next page

Table 11 (continued)

Dataset	Method	Clean	PGD	CE	T	Square	ℓ_2	WCRA
StanfordCars	Zero-shot	65.24	0.00	0.00	0.00	0.07	0.00	0.00
	TPT	66.80	0.00	5.02	5.06	51.13	0.00	0.00
	R-TPT	67.07	23.99	38.13	36.21	57.06	44.21	19.85
	DBD	**68.08**	**88.31**	**69.46**	**65.51**	**61.51**	**92.50**	**55.03**
	MTA	67.96	11.85	17.82	24.21	50.86	24.46	9.71
	TTC	56.80	3.27	6.08	9.65	49.25	0.07	0.07
	MAC	64.69	49.26	55.04	25.21	51.80	65.79	17.52
Food101	Zero-shot	83.13	0.00	0.00	0.00	0.24	0.00	0.00
	TPT	84.21	0.00	2.73	2.73	71.56	0.00	0.00
	R-TPT	83.84	32.22	45.54	52.69	75.93	64.45	28.75
	DBD	83.81	**94.30**	**85.96**	**80.10**	**76.15**	**99.49**	**72.21**
	MTA	**84.59**	19.02	28.76	44.04	72.96	48.26	16.77
	TTC	78.62	3.72	3.90	15.14	64.67	0.37	0.08
	MAC	82.22	59.79	76.36	45.05	69.15	85.67	32.87
SUN397	Zero-shot	62.23	0.00	0.00	0.00	0.38	0.00	0.00
	TPT	64.15	0.00	5.72	5.84	55.30	0.04	0.00
	R-TPT	64.23	34.94	45.75	47.02	58.61	51.86	31.76
	DBD	64.13	**91.99**	**67.13**	**64.45**	**60.25**	**97.48**	**57.91**
	MTA	**64.26**	18.86	25.02	34.89	54.34	31.68	16.78
	TTC	52.64	3.00	5.88	12.46	48.70	1.17	0.12
	MAC	62.31	48.24	56.31	34.83	52.53	69.52	25.00
EuroSAT	Zero-shot	52.06	0.00	0.00	0.00	0.00	0.00	0.00
	TPT	**54.59**	0.00	9.22	9.21	32.28	0.00	0.00
	R-TPT	45.83	5.89	13.93	19.51	33.38	35.04	2.20
	DBD	38.75	**67.43**	**56.19**	**40.84**	31.11	**95.67**	**22.80**
	MTA	51.44	0.64	2.31	5.49	24.73	9.73	0.15
	TTC	54.06	5.40	9.09	15.54	**36.80**	2.33	0.38
	MAC	53.15	21.16	30.47	21.98	30.54	48.10	5.80
ImageNet	Zero-shot	66.77	0.00	0.00	0.00	0.57	0.00	0.00
	TPT	69.06	0.00	4.88	4.94	59.00	0.02	0.00
	R-TPT	**69.37**	31.35	43.44	46.00	62.92	50.07	27.98
	DBD	68.54	**91.21**	**70.46**	**67.47**	**63.43**	**96.85**	**60.15**
	MTA	69.35	21.36	28.08	38.11	58.22	34.99	19.14
	TTC	53.56	4.72	5.68	12.52	50.65	1.48	0.21
	MAC	66.63	47.75	58.89	34.74	55.53	68.26	24.97
ImageNetV2	Zero-shot	60.69	0.01	0.00	0.00	0.49	0.00	0.00
	TPT	62.83	0.02	4.95	4.99	52.81	0.00	0.00
	R-TPT	**63.61**	26.77	38.54	40.67	57.02	44.23	23.61
	DBD	62.53	**88.95**	**64.63**	**61.27**	**57.16**	**96.08**	**53.64**
	MTA	63.34	17.11	23.14	32.56	52.09	28.78	15.02
	TTC	49.23	3.48	5.26	10.19	43.21	1.33	0.18
	MAC	59.94	41.34	52.35	29.95	49.26	61.90	20.40
ImageNetA	Zero-shot	47.85	0.01	0.00	0.00	0.07	0.00	0.00
	TPT	54.52	0.01	7.87	7.87	39.76	0.01	0.00
	R-TPT	56.97	13.95	28.56	32.11	**46.05**	26.24	11.52
	DBD	49.52	**82.59**	**53.81**	**45.29**	39.08	**97.36**	**34.64**
	MTA	**57.01**	7.92	17.45	25.51	41.89	13.28	6.65
	TTC	43.81	1.53	5.56	8.53	30.99	0.21	0.03
	MAC	44.52	21.23	30.75	17.61	29.04	40.51	7.77

Continued on next page

Table 11 (continued)

Dataset	Method	Clean	PGD	CE	T	Square	ℓ_2	WCRA
ImageNetR	Zero-shot	73.84	0.13	0.00	0.00	2.67	0.09	0.00
	TPT	**77.01**	0.32	4.87	5.09	67.47	0.30	0.00
	R-TPT	**77.01**	46.25	55.71	59.24	70.59	63.23	43.03
	DBD	76.65	**94.90**	**78.46**	**75.12**	**73.28**	**98.66**	**70.06**
	MTA	76.95	32.32	38.27	50.67	67.04	46.21	30.21
	TTC	65.78	10.96	5.32	17.61	61.39	8.73	0.36
	MAC	73.98	53.41	68.95	37.87	64.91	74.31	28.59
ImageNetSketch	Zero-shot	45.97	0.06	0.00	0.00	0.71	0.02	0.00
	TPT	47.80	0.23	4.58	4.81	37.56	0.10	0.00
	R-TPT	47.56	25.84	31.37	33.66	39.74	36.50	22.52
	DBD	48.31	**83.97**	**49.96**	**47.19**	**43.94**	**94.46**	**41.23**
	MTA	**48.41**	13.97	15.14	24.56	36.79	18.96	11.02
	TTC	39.42	9.57	4.22	14.47	35.94	9.03	0.25
	MAC	45.94	29.54	41.70	18.06	36.89	49.12	11.45
DataComp-B/16								
DTD	Zero-shot	56.21	0.12	0.00	0.00	4.02	0.53	0.00
	TPT	**57.21**	0.30	4.85	4.96	48.35	1.12	0.00
	R-TPT	56.97	40.66	46.51	48.05	50.89	51.36	37.77
	DBD	57.03	**87.77**	**60.76**	**58.63**	**52.07**	**94.62**	**50.89**
	MTA	56.74	27.13	28.37	35.99	45.69	35.46	25.06
	TTC	53.31	4.20	7.57	15.78	43.20	7.62	0.83
	MAC	54.14	49.11	52.72	39.07	44.74	63.89	30.67
Caltech101	Zero-shot	97.04	0.37	0.00	0.00	19.51	0.53	0.00
	TPT	97.12	0.65	0.97	1.05	93.35	2.23	0.00
	R-TPT	96.47	85.76	88.84	90.99	94.16	93.79	83.94
	DBD	**97.77**	**99.19**	**97.73**	**97.24**	**96.47**	**99.88**	**95.86**
	MTA	97.57	77.73	83.37	89.17	94.93	90.55	76.51
	TTC	96.39	36.23	5.88	42.96	89.61	26.49	1.58
	MAC	96.02	91.28	96.43	89.82	94.16	96.35	83.20
Flowers102	Zero-shot	75.64	0.00	0.00	0.00	2.23	0.00	0.00
	TPT	75.31	0.00	2.60	2.68	69.31	0.00	0.00
	R-TPT	75.92	49.94	59.48	62.16	72.63	67.11	47.71
	DBD	**76.13**	**90.34**	**78.73**	**77.63**	**74.10**	**98.17**	**72.31**
	MTA	**76.13**	35.24	43.44	54.97	70.40	58.42	33.58
	TTC	74.87	3.33	5.32	25.62	64.31	1.91	0.37
	MAC	76.00	71.21	74.62	45.68	69.83	81.73	39.59
OxfordPets	Zero-shot	90.87	0.00	0.00	0.00	1.09	0.00	0.00
	TPT	90.90	0.00	3.30	3.30	68.60	0.00	0.00
	R-TPT	90.95	55.52	68.38	70.81	84.71	79.37	50.72
	DBD	**91.71**	**98.69**	**93.46**	**92.15**	**88.93**	**99.24**	**87.08**
	MTA	**91.71**	42.68	56.66	64.57	82.99	69.80	40.20
	TTC	72.77	5.04	5.86	24.37	73.78	2.21	0.27
	MAC	91.58	84.57	91.03	62.99	82.39	91.14	54.76
FGVCAircraft	Zero-shot	29.46	0.00	0.00	0.00	0.00	0.00	0.00
	TPT	32.28	0.00	7.08	7.08	24.69	0.00	0.00
	R-TPT	32.64	13.29	21.06	21.54	25.56	20.79	11.07
	DBD	**33.84**	**69.94**	**33.66**	**30.45**	**27.81**	**73.60**	**22.71**
	MTA	33.48	5.19	9.93	13.95	21.54	11.34	4.17
	TTC	26.13	1.80	3.81	6.78	18.21	0.00	0.03
	MAC	28.29	31.14	27.93	17.07	22.02	42.66	10.62

Continued on next page

Table 11 (continued)

Dataset	Method	Clean	PGD	CE	T	Square	ℓ_2	WCRA
UCF101	Zero-shot	62.04	0.00	0.00	0.00	0.24	0.00	0.00
	TPT	**63.97**	0.00	4.73	4.73	55.75	0.00	0.00
	R-TPT	63.36	35.03	44.38	47.16	58.18	50.75	32.46
	DBD	63.12	**92.10**	**66.06**	**62.83**	**59.00**	**98.55**	**57.02**
	MTA	63.60	24.21	32.91	41.92	57.65	41.45	22.34
	TTC	62.09	2.54	4.07	16.07	50.22	1.11	0.05
	MAC	61.41	53.16	58.68	41.00	52.63	71.93	32.09
StanfordCars	Zero-shot	89.96	0.00	0.01	0.01	4.14	0.00	0.00
	TPT	90.71	0.00	1.92	1.93	77.53	0.00	0.00
	R-TPT	90.65	53.82	64.93	66.63	83.02	75.71	50.06
	DBD	91.44	**98.66**	**91.99**	**89.63**	**86.34**	**99.03**	**84.32**
	MTA	**91.58**	41.55	53.67	61.42	82.80	64.73	38.94
	TTC	85.34	12.20	4.17	19.90	72.47	0.56	0.31
	MAC	88.05	86.73	87.76	68.14	79.53	91.95	60.15
Food101	Zero-shot	87.16	0.00	0.00	0.00	2.83	0.00	0.00
	TPT	88.14	0.00	2.40	2.38	76.67	0.00	0.00
	R-TPT	87.67	47.70	59.20	67.13	81.48	73.34	44.66
	DBD	87.88	**97.23**	**89.81**	**87.12**	**82.72**	**99.83**	**80.92**
	MTA	**88.62**	36.50	48.27	63.06	81.12	65.00	33.97
	TTC	85.52	11.26	3.95	23.37	62.35	0.37	0.35
	MAC	86.92	78.78	85.69	68.87	76.44	93.24	57.99
SUN397	Zero-shot	68.90	0.00	0.01	0.00	0.88	0.01	0.00
	TPT	69.66	0.00	4.59	4.61	59.68	0.02	0.00
	R-TPT	69.62	42.66	52.78	54.51	63.63	58.20	39.94
	DBD	**70.58**	**96.32**	**73.48**	**71.34**	**65.13**	**98.92**	**63.93**
	MTA	70.50	31.62	40.96	48.58	61.43	46.31	29.98
	TTC	56.48	6.28	5.23	15.51	50.48	1.74	0.29
	MAC	68.15	66.95	67.84	51.35	58.99	80.79	41.39
EuroSAT	Zero-shot	51.48	0.00	0.05	0.05	0.00	0.00	0.00
	TPT	51.38	0.00	6.85	7.05	29.19	0.00	0.00
	R-TPT	46.70	5.23	15.48	22.64	34.64	29.48	2.99
	DBD	41.26	**71.23**	**58.06**	**49.41**	**35.48**	**97.07**	**27.89**
	MTA	51.59	1.42	5.93	10.58	29.58	13.32	0.70
	TTC	**55.48**	6.63	11.96	17.75	34.91	4.42	0.53
	MAC	42.21	27.17	37.95	26.84	27.77	55.40	9.70
ImageNet	Zero-shot	73.31	0.00	0.00	0.00	1.53	0.00	0.00
	TPT	74.89	0.00	3.75	3.77	65.01	0.04	0.00
	R-TPT	75.24	42.65	53.78	56.83	69.24	59.86	39.90
	DBD	74.60	**94.40**	**76.73**	**74.52**	**69.68**	**98.48**	**67.62**
	MTA	**75.33**	36.09	45.32	53.64	66.39	52.12	34.19
	TTC	65.01	8.92	4.58	17.27	53.15	2.69	0.35
	MAC	72.60	68.08	71.77	53.09	62.53	81.17	43.67
ImageNetV2	Zero-shot	65.36	0.00	0.00	0.00	1.16	0.01	0.00
	TPT	67.30	0.00	4.19	4.21	57.27	0.06	0.00
	R-TPT	**67.90**	35.75	46.76	50.07	61.47	51.33	33.33
	DBD	67.08	**92.82**	**69.44**	**66.55**	**62.01**	**97.92**	**59.82**
	MTA	67.80	29.26	38.58	46.61	58.74	43.46	27.76
	TTC	47.96	6.93	4.64	15.73	42.95	2.75	0.31
	MAC	64.25	60.79	63.18	45.84	54.66	75.11	36.66

Continued on next page

Table 11 (continued)

Dataset	Method	Clean	PGD	CE	T	Square	ℓ_2	WCRA
ImageNetA	Zero-shot	47.67	0.00	0.00	0.00	0.05	0.00	0.00
	TPT	54.39	0.01	7.69	7.59	38.41	0.00	0.00
	R-TPT	58.36	16.24	32.36	36.65	**48.07**	29.92	14.07
	DBD	50.45	**86.80**	**54.72**	**46.75**	38.97	**98.61**	**35.61**
	MTA	**58.69**	10.95	23.79	32.23	44.49	20.57	9.61
	TTC	42.84	2.31	3.63	7.35	25.53	0.37	0.05
	MAC	44.39	32.39	39.55	23.40	28.53	53.99	12.23
ImageNetR	Zero-shot	82.14	0.16	0.01	0.00	10.17	0.14	0.00
	TPT	84.04	0.38	3.35	3.39	75.12	0.30	0.00
	R-TPT	84.54	58.38	66.01	70.07	78.40	73.09	55.53
	DBD	**84.71**	**96.93**	**86.00**	**82.83**	**81.12**	**99.55**	**78.57**
	MTA	84.59	49.70	56.26	66.64	76.45	64.55	47.87
	TTC	78.01	17.04	4.02	22.79	68.21	15.25	0.50
	MAC	81.40	75.67	81.95	62.02	74.37	89.37	53.44
ImageNetSketch	Zero-shot	59.22	0.16	0.00	0.00	3.46	0.07	0.00
	TPT	60.49	0.48	3.91	4.10	49.87	0.24	0.00
	R-TPT	60.40	39.14	43.83	47.86	52.17	50.82	35.73
	DBD	**61.34**	**86.78**	**63.52**	**59.85**	**56.40**	**97.07**	**53.34**
	MTA	61.05	31.25	34.14	43.28	49.16	41.24	28.96
	TTC	56.76	16.40	2.72	18.63	47.40	22.93	0.36
	MAC	57.30	48.16	58.15	38.16	48.85	66.92	28.11
SigLIP-B/16								
DTD	Zero-shot	62.94	0.24	0.06	0.00	8.81	1.48	0.00
	TPT	63.12	0.35	1.12	1.06	50.24	1.65	0.00
	R-TPT	62.83	48.23	53.37	54.31	57.15	58.04	46.10
	DBD	62.65	**89.78**	**65.31**	**63.53**	**57.86**	**97.22**	**56.56**
	MTA	**63.30**	36.47	36.05	45.63	54.55	42.85	32.33
	TTC	55.56	6.03	7.57	17.61	48.05	15.37	1.36
	MAC	62.12	45.27	53.19	39.66	52.90	53.31	29.55
Caltech101	Zero-shot	96.84	1.54	0.08	0.04	35.82	6.25	0.04
	TPT	97.12	2.64	0.49	1.62	91.85	7.91	0.04
	R-TPT	96.63	89.21	92.01	93.02	94.97	94.52	87.95
	DBD	96.15	**99.39**	**97.40**	**97.16**	**96.39**	**99.92**	**95.66**
	MTA	**97.16**	83.16	86.82	91.97	95.21	91.76	81.78
	TTC	96.11	26.57	9.01	44.02	88.15	29.29	3.04
	MAC	96.67	83.81	93.96	87.30	94.56	90.99	77.12
Flowers102	Zero-shot	85.71	0.00	0.00	0.00	6.98	0.00	0.00
	TPT	86.11	0.00	1.02	1.02	76.53	0.00	0.00
	R-TPT	**87.58**	66.46	73.77	76.21	82.01	78.93	64.60
	DBD	87.29	**93.79**	**86.56**	**86.52**	**83.15**	**97.89**	**80.43**
	MTA	87.13	54.61	60.78	71.50	80.80	68.90	51.89
	TTC	71.13	15.31	7.92	24.48	70.36	3.82	1.58
	MAC	86.24	60.94	81.49	44.42	80.88	70.36	34.19
OxfordPets	Zero-shot	89.81	0.00	0.00	0.00	5.01	0.03	0.00
	TPT	89.97	0.00	0.60	0.60	75.36	0.03	0.00
	R-TPT	**90.02**	58.14	68.55	71.52	83.78	75.52	55.98
	DBD	88.85	**97.82**	**90.49**	**89.89**	**84.76**	**98.91**	**82.53**
	MTA	89.72	47.10	56.39	67.87	82.94	66.45	44.64
	TTC	68.14	5.34	6.84	22.19	70.78	6.11	0.57
	MAC	89.40	69.47	84.27	52.11	79.45	76.26	41.37

Continued on next page

Table 11 (continued)

Dataset	Method	Clean	PGD	CE	T	Square	ℓ_2	WCRA
FGVCAircraft	Zero-shot	43.92	0.00	0.00	0.06	0.06	0.00	0.00
	TPT	**44.61**	0.00	1.59	1.62	24.54	0.00	0.00
	R-TPT	34.53	21.15	29.16	28.98	36.06	30.00	18.81
	DBD	25.65	**71.74**	**47.55**	**42.78**	**41.19**	**74.95**	**33.03**
	MTA	44.28	9.51	13.77	19.80	32.52	15.18	7.86
	TTC	28.29	0.24	5.79	9.60	26.58	0.27	0.00
	MAC	44.01	39.06	40.02	20.10	34.02	47.55	14.25
UCF101	Zero-shot	60.32	0.00	0.00	0.00	0.95	0.00	0.00
	TPT	61.41	0.00	1.48	1.48	46.63	0.00	0.00
	R-TPT	59.42	35.79	43.75	47.58	53.79	47.79	34.05
	DBD	58.63	**83.11**	**63.68**	**60.16**	**55.46**	**96.30**	**53.13**
	MTA	**61.59**	26.67	34.47	44.17	53.71	37.77	24.77
	TTC	59.00	2.67	3.57	13.77	47.16	0.82	0.16
	MAC	59.77	41.37	50.83	35.29	50.67	55.70	25.64
StanfordCars	Zero-shot	90.70	0.00	0.00	0.00	7.81	0.10	0.00
	TPT	91.26	0.00	0.93	0.93	81.88	0.10	0.00
	R-TPT	90.56	67.98	77.35	78.19	87.97	80.62	65.90
	DBD	87.87	**97.67**	**91.89**	**90.92**	**88.47**	**98.63**	**86.34**
	MTA	**91.61**	58.00	68.85	74.21	86.81	71.32	55.81
	TTC	76.57	8.69	8.06	19.76	74.74	3.20	1.03
	MAC	90.23	80.59	87.23	56.14	83.66	77.23	50.07
Food101	Zero-shot	86.82	0.00	0.00	0.00	2.00	0.01	0.00
	TPT	87.17	0.00	0.83	0.84	66.24	0.01	0.00
	R-TPT	86.75	53.10	63.77	68.47	75.57	72.02	50.02
	DBD	87.21	**94.92**	**88.65**	**85.25**	**77.85**	**99.25**	**75.43**
	MTA	**87.86**	42.85	54.78	66.13	76.41	62.63	40.25
	TTC	43.87	8.24	4.49	20.62	49.03	2.53	0.38
	MAC	86.18	61.47	74.02	58.66	72.41	78.67	42.71
SUN397	Zero-shot	67.42	0.00	0.00	0.00	3.51	0.04	0.00
	TPT	67.97	0.00	1.67	1.66	53.86	0.05	0.00
	R-TPT	67.18	44.91	53.95	54.46	61.64	57.34	42.93
	DBD	67.13	**92.57**	**71.27**	**69.14**	**63.48**	**97.59**	**61.45**
	MTA	**68.16**	35.06	43.13	50.74	60.86	45.45	33.01
	TTC	49.57	5.93	5.98	15.30	48.59	4.21	0.35
	MAC	67.09	50.75	56.65	40.87	57.28	63.46	29.96
EuroSAT	Zero-shot	**43.79**	0.00	0.00	0.00	0.00	0.00	0.00
	TPT	43.17	0.00	1.51	1.48	11.33	0.00	0.00
	R-TPT	29.78	4.98	7.72	14.67	15.19	22.67	3.01
	DBD	29.84	**53.38**	**46.27**	**36.98**	18.00	**93.54**	**16.54**
	MTA	37.21	1.83	3.57	9.75	13.19	9.26	0.91
	TTC	39.32	4.30	6.04	10.05	**20.28**	4.40	0.21
	MAC	40.00	18.62	25.21	18.19	16.85	32.09	6.42
ImageNet	Zero-shot	75.93	0.00	0.00	0.00	4.85	0.09	0.00
	TPT	76.62	0.01	1.17	1.17	61.63	0.11	0.00
	R-TPT	77.13	49.59	59.66	61.78	71.03	62.62	47.19
	DBD	76.47	**92.37**	**78.04**	**76.18**	**71.08**	**97.55**	**68.15**
	MTA	**77.93**	44.37	52.62	60.48	69.67	55.00	42.28
	TTC	47.89	9.18	5.88	19.14	52.63	6.97	0.62
	MAC	75.31	56.01	66.14	48.09	64.63	65.21	36.34

Continued on next page

Table 11 (continued)

Dataset	Method	Clean	PGD	CE	T	Square	ℓ_2	WCRA
ImageNetV2	Zero-shot	68.41	0.00	0.00	0.00	4.44	0.08	0.00
	TPT	69.09	0.00	1.26	1.24	53.86	0.12	0.00
	R-TPT	69.82	41.59	51.64	53.56	**63.65**	53.96	39.35
	DBD	68.53	**90.01**	**70.89**	**68.39**	63.16	**96.82**	**60.17**
	MTA	**70.77**	36.05	44.59	52.05	61.88	45.94	34.25
	TTC	45.55	6.97	5.44	15.76	47.22	5.60	0.44
	MAC	67.39	48.95	57.99	40.25	56.64	59.79	29.34
ImageNetA	Zero-shot	44.99	0.00	0.01	0.01	0.72	0.05	0.00
	TPT	46.59	0.00	1.72	1.75	22.72	0.05	0.00
	R-TPT	53.81	16.40	28.69	32.41	38.53	26.60	14.35
	DBD	49.80	**76.95**	**50.64**	**42.05**	33.93	**96.11**	**30.25**
	MTA	**57.95**	11.87	23.20	32.15	**39.00**	19.11	10.37
	TTC	30.76	1.60	3.79	7.52	24.37	1.56	0.09
	MAC	40.35	20.27	26.67	16.64	25.96	31.73	8.21
ImageNetR	Zero-shot	89.32	1.27	0.08	0.02	15.66	3.99	0.00
	TPT	89.80	1.40	0.88	0.83	78.92	4.27	0.00
	R-TPT	89.29	72.54	77.93	80.88	85.88	82.57	70.56
	DBD	89.06	**96.92**	**91.72**	**89.45**	**87.87**	**99.22**	**85.93**
	MTA	**90.83**	65.93	71.32	79.96	85.82	75.67	64.01
	TTC	72.05	22.21	6.56	29.55	74.45	17.95	1.65
	MAC	89.16	70.88	84.21	64.32	82.48	79.92	53.31
ImageNetSketch	Zero-shot	66.58	1.36	0.04	0.00	8.17	3.62	0.00
	TPT	67.15	1.61	1.40	1.38	52.40	4.17	0.00
	R-TPT	65.65	48.65	52.76	55.66	58.74	58.02	46.09
	DBD	64.32	**85.82**	**68.53**	**65.47**	**62.45**	**93.69**	**58.68**
	MTA	**67.56**	40.81	43.23	52.03	56.64	48.69	38.50
	TTC	55.12	23.84	4.05	27.01	54.62	19.91	0.79
	MAC	65.24	45.71	61.25	40.10	56.14	56.67	30.19

E ADDITIONAL BACKBONES UNDER A COMMON TWO-ATTACK COMPARISON

The six local datasets also have complete clean, PGD-10, and APGD-CE results for CLIP-B/32 and CLIP-L/14. Together with the three core backbones this forms 630 evaluations, including 252 runs outside the primary three-backbone grid. Table 12 reports the same two-attack mean for every backbone, without treating it as four-attack WCRA or silently adding missing attacks.

Table 12: Five-backbone comparison over the same six datasets. Each cell is clean accuracy / mean accuracy over PGD-10 and APGD-CE (%). Best and second best are marked separately for each number within each backbone.

Method	CLIP-B/32	CLIP-B/16	CLIP-L/14	DataComp	SigLIP
Zero-shot	60.4 / 0.1	63.9 / 0.0	70.3 / 0.1	68.5 / 0.0	73.3 / 0.2
TPT	60.2 / 2.0	64.8 / 1.8	71.2 / 2.3	69.5 / 2.0	73.7 / 0.8
R-TPT	60.2 / 38.3	64.9 / 41.8	71.2 / 50.8	69.4 / 50.7	71.8 / 56.6
DBD	**61.7** / **67.1**	**65.1** / **75.2**	71.2 / **80.9**	**69.9** / **80.7**	69.9 / **82.2**
MTA	61.6 / 25.9	64.9 / 24.9	**71.5** / 34.8	**69.9** / 38.9	**73.9** / 45.5
TTC	58.1 / 5.6	59.9 / 5.3	68.6 / 8.0	64.3 / 7.1	63.0 / 8.1
MAC	60.2 / 35.5	64.2 / 50.9	70.2 / 58.4	67.9 / 65.2	73.0 / 62.0

F ALL-TASK VIEW OF THE CORE SUITE

(a) CLIP-B/16

	DTD	Caltech	Flowers	Pets	Aircraft	UCF	Cars	Food	SUN	EuroSAT	IN	IN-V2	IN-A	IN-R	Sketch
Zero-shot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TPT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R-TPT	25	69	33	38	8	32	20	29	32	2	28	24	12	43	23
DBD	38	90	63	80	18	61	55	72	58	23	60	54	35	70	41
MTA	11	49	15	25	2	19	10	17	17	0	19	15	7	30	11
TTC	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
MAC	13	56	15	28	5	29	18	33	25	6	25	20	8	29	11

(b) DataComp-B/16

	DTD	Caltech	Flowers	Pets	Aircraft	UCF	Cars	Food	SUN	EuroSAT	IN	IN-V2	IN-A	IN-R	Sketch
Zero-shot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TPT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R-TPT	38	84	48	51	11	32	50	45	40	3	40	33	14	56	36
DBD	51	96	72	87	23	57	84	81	64	28	68	60	36	79	53
MTA	25	77	34	40	4	22	39	34	30	1	34	28	10	48	29
TTC	1	2	0	0	0	0	0	0	0	1	0	0	0	0	0
MAC	31	83	40	55	11	32	60	58	41	10	44	37	12	53	28

(c) SigLIP-B/16

	DTD	Caltech	Flowers	Pets	Aircraft	UCF	Cars	Food	SUN	EuroSAT	IN	IN-V2	IN-A	IN-R	Sketch
Zero-shot	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
TPT	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R-TPT	46	88	65	56	19	34	66	50	43	3	47	39	14	71	46
DBD	57	96	80	83	33	53	86	75	61	17	68	60	30	86	59
MTA	32	82	52	45	8	25	56	40	33	1	42	34	10	64	38
TTC	1	3	2	1	0	0	1	0	0	0	1	0	0	2	1
MAC	30	77	34	41	14	26	50	43	30	6	36	29	8	53	30

Figure 7: All 315 primary family WCRA values, rounded to integer percentages; color scale 0–100. Individual accuracies appear in Table 11. Dataset labels: DTD; Caltech (Caltech101); Flowers (Flowers102); Pets (OxfordPets); Aircraft (FGVCAircraft); UCF (UCF101); Cars (StanfordCars); Food (Food101); SUN (SUN397); EuroSAT; IN (ImageNet); IN-V2 (ImageNetV2); IN-A (ImageNet-A); IN-R (ImageNet-R); Sketch (ImageNetSketch). Exact WCRA values are retained in the exported family table.

G ADAPTIVE RESULTS AND FULL-METHOD ATTACK PROTOCOL

Section 6 reports the 100-image study and matched ten-image budget comparison; the tables below give clean-correct denominators and results for each method. The separately reported eight-image pilot overlaps the main cohort and supplies no independent evidence.

100-image results. Table 13 reports accuracy and clean-correct loss for all seven methods under the same query budget.

Table 13: 100-image DTD results on CLIP-B/16: 100 candidates and four search draws, with 16 held-out draws per image. CC-loss counts are broken clean-correct draws / total clean-correct draws. Ranks apply to the seven-method accuracy columns.

Method	Clean (%)	Cached PGD-10 (%)	Adaptive (%)	CC-loss counts
Zero-shot	42.0000	0.0000	0.0000	672/672
TPT	42.6875	0.0000	0.0000	683/683
R-TPT	46.0625	31.1250	29.4375	273/737
DBD	40.0000	75.5625	34.3750	95/640
MTA	41.8750	14.1250	14.2500	442/670
TTC	42.3125	4.9375	3.6250	619/677
MAC	41.1250	26.8750	23.5625	296/658

Table 14: Matched ten-image DTD comparison. Entries are accuracy (%). Clean/cached held-out predictions are identical across budgets. Comparisons use the same ten IDs; bold/shading and underline mark the best and second distinct values.

Method	Clean	Cached	100 candidates	500 candidates
Zero-shot	20.00	0.00	0.00	0.00
TPT	10.00	0.00	0.00	0.00
R-TPT	10.00	8.75	**9.38**	**9.38**
DBD	13.75	**78.12**	8.75	2.50
MTA	20.00	0.00	0.00	0.00
TTC	19.38	0.00	0.00	0.00
MAC	**22.50**	7.50	6.25	3.75

The 500-candidate configuration lowers matched-subset accuracy relative to 100 candidates from 8.75% to 2.50% for DBD and from 6.25% to 3.75% for MAC. R-TPT remains at 9.38%; the other four methods remain at zero. Among paired clean-correct draws, DBD loses 8/22 at 100 candidates and 18/22 at 500; MAC loses 32/36 and 34/36, respectively. Appendix H reports paired uncertainty and shows why this small budget comparison does not establish attack sufficiency.

All-method eight-image pilot. The pilot checks the evaluation procedure, but its robustness estimate is not representative of the population.

Table 15: Eight-image adaptive pilot on CLIP-B/16. Accuracy averages correctness over 16 held-out method draws per image, then over images. The final column counts attacked errors among paired clean-correct draws.

Method	Clean	Cached PGD-10	Adaptive	Clean-correct draws lost
Zero-shot	12.50	0.00	0.00	16/16
TPT	12.50	0.00	0.00	16/16
R-TPT	12.50	10.94	11.72	1/16
DBD	11.72	72.66	10.94	1/15
MTA	12.50	0.00	0.00	16/16
TTC	13.28	0.00	0.00	17/17
MAC	16.41	5.47	2.34	21/21

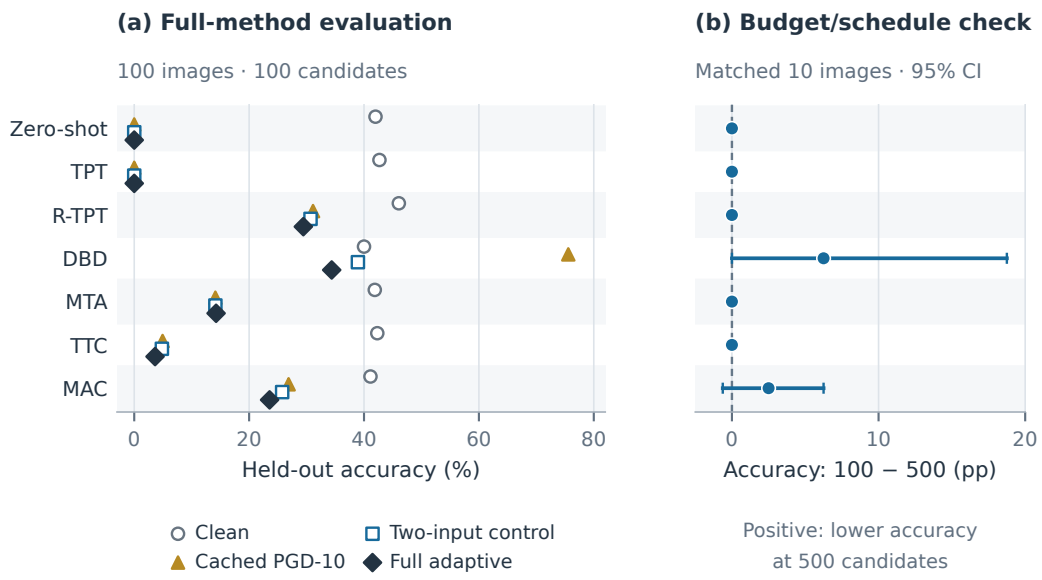

Figure 8: Adaptive results on CLIP-B/16 and DTD. Left: the 100-image cohort under clean inputs, cached PGD-10, the two-input initialization control, and full adaptive search. The control selects clean or cached inputs using search margins. Right: paired accuracy decrease from 100 to 500 candidates on the same ten images, with conditional 95% image-bootstrap intervals. Positive values indicate lower accuracy at 500 candidates; neither DBD nor MAC excludes zero. Accuracies average sixteen held-out draws per image. The right panel is a budget-and-schedule diagnostic on a smaller cohort.

DBD has low clean accuracy on these eight images, and only one of its 15 clean-correct draws becomes incorrect. Its large cached-to-adaptive reduction chiefly removes the apparent gain on cached PGD-10 inputs; it does not establish broad adaptive collapse. R-TPT's adaptive accuracy slightly exceeds its cached value under held-out randomness, which is possible because held-out correctness is not the search objective.

Attacker access and input domain. Each score query runs the complete method: preprocessing, view generation, adaptation, and prediction. The attacker observes class scores and knows the true label; the defense sees only the candidate image and common model/prompt context. The attack does not differentiate through the adaptation optimizer. Inputs are canonical 224×224 RGB images quantized to uint8, with an ℓ_∞ radius of four intensity levels. Every query resets episodic state.

Search and randomization. Square-style patch proposals (Andriushchenko et al., 2020) minimize the mean true-versus-best-other logit margin over four fixed method seeds. The first three candidates are the clean image, matching cached PGD-10 image, and random vertical stripes; subsequent candidates modify the best image under this objective. Budgets of 100 and 500 candidates cost 400 and 2,000 full-method calls per image, including initialization but excluding PGD-10 warm-start generation. Neither early stopping nor correctness-based filtering is used.

Sixteen disjoint held-out method seeds score the selected, clean, and cached PGD-10 images, requiring 48 additional calls per image. The same bank supports paired comparisons across these conditions and never selects candidates. Held-out scoring checks overfitting to the fixed search bank; more candidates alone neither establish convergence nor remove this risk.

Shared cohort and statistics. All seven methods use the same 100 DTD images with CLIP-B/16, selected independently of model predictions by seeded round-robin over shuffled classes and examples. The stronger budget uses the first ten IDs, compared against those same images at the smaller budget, not the 100-image mean.

Accuracy averages correctness over 16 individual held-out predictions per image, then over images; it is not ensemble accuracy from mean logits. For each method, we also report the fraction of paired

clean-correct draws made incorrect by the selected attack, stating both numerator and denominator. Draws within images are correlated and count as neither additional test images nor independent benchmark seed replications. The two budgets use different proposal schedules, so search trajectories need not be nested, and held-out scores need not decrease monotonically.

Interpretation. This bounded full-method attack does not replace broader adaptive evaluation across models, datasets, restart budgets, objectives, and gradient-based strategies. Failure to find an error within its budget does not establish robustness to every admissible perturbation.

H ADAPTIVE UNCERTAINTY AND INITIALIZATION CONTROL

Paired image uncertainty. Table 16 reports accuracy differences with 95% percentile intervals from 10,000 paired bootstrap samples of the same 100 image IDs (seed 20260926). Each image retains all sixteen held-out draws, and the same sampled IDs are used for all methods and contrasts. These intervals condition on this cohort, the selected attacks, and the fixed search/evaluation seed banks; they do not measure variation over new attack seeds or datasets. Draws within an image are not independent sample units. Positive differences indicate lower accuracy after the full search; intervals are pointwise, without multiplicity correction.

Table 16: Accuracy differences in percentage points, with paired image-bootstrap 95% intervals on the 100-image DTD cohort. The control selects between clean and cached PGD-10 using search draws only, then uses held-out predictions. It does not include the random-stripe initial candidate.

Method	Clean − adaptive	Cached − adaptive	Control − adaptive
Zero-shot	42.00 [33.00, 52.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
TPT	42.69 [33.69, 52.19]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
R-TPT	16.62 [10.62, 23.25]	1.69 [0.31, 3.38]	1.25 [0.06, 2.81]
DBD	5.62 [2.00, 10.06]	41.19 [32.00, 50.38]	4.56 [1.44, 8.50]
MTA	27.62 [19.25, 36.44]	-0.12 [-0.81, 0.50]	-0.12 [-0.81, 0.50]
TTC	38.69 [30.25, 47.75]	1.31 [0.19, 2.81]	1.19 [0.06, 2.69]
MAC	17.56 [11.31, 24.13]	3.31 [0.31, 6.69]	2.19 [-0.25, 5.06]

What the initialization control tests. For each image and method, we select the lower saved mean search-bank margin among the clean and cached PGD-10 candidates, with ties retaining clean. We then report the saved held-out predictions for that selected input. Neither held-out scores nor held-out correctness select the control. This separates comparison to an attack that can choose either existing input from comparison to the cached input alone. It is a two-candidate control, not the complete three-candidate initialization: independent held-out predictions for the random-stripe candidate were not saved.

For DBD, the control selects clean on 92 of 100 images and cached PGD-10 on 8. Its held-out accuracy is 38.94%, compared with 34.38% after the full search. The much larger cached-to-adaptive difference therefore includes removing gains specific to cached inputs.

Search progress and its limits. The saved traces record candidate margins and strict improvements on the four search draws. Comparing the best margin after candidate three with the final margin measures optimization beyond all three initial candidates. The final accepted candidate identifies whether the output was clean, cached, random-stripe, or a later query. These traces measure the search objective; intermediate stripe and patch candidates have no saved held-out predictions. The traces therefore cannot supply a held-out convergence curve or establish optimality, and logit-margin reductions are not normalized for comparison across methods. For DBD, the search improves the margin beyond the first three candidates on 100 of 100 images; 100 final images come from later queries. Objective improvement alone does not guarantee a held-out accuracy decrease. **Budget-check cohort.** The 100-image cohort covers all 47 classes, with 2–3 images per class. The first ten IDs cover 10 classes with one image each. Across completed methods, their clean accuracies are 10.00–22.50%, compared with 40.00–46.06% on the full 100 images. Their lower clean accuracy and limited class coverage make the stronger-budget comparison a diagnostic on matched examples, not a representative robustness estimate.

Completed matched-budget comparison. All seven methods have both budgets on the same ten IDs, with identical clean/cached held-out predictions and the same evaluation seeds. Table 17 reports accuracy changes and paired clean-correct loss counts. Positive differences mean lower accuracy at 500 candidates; the image-bootstrap intervals use the same procedure as above, now resampling ten images and retaining all sixteen draws.

Table 17: Matched ten-image budget effects. CC-loss entries are broken clean-correct draws / all clean-correct draws. Accuracy differences (percentage points) carry conditional 95% image-bootstrap intervals. Counts from different methods have different clean-correct denominators; they are not ranked. These ten images are a budget diagnostic, not an independent test set.

Method	CC-loss, 100	CC-loss, 500	Accuracy $100 - 500$ (pp)
Zero-shot	32/32	32/32	0.00 [0.00, 0.00]
TPT	16/16	16/16	0.00 [0.00, 0.00]
R-TPT	1/16	1/16	0.00 [0.00, 0.00]
DBD	8/22	18/22	6.25 [0.00, 18.75]
MTA	32/32	32/32	0.00 [0.00, 0.00]
TTC	31/31	31/31	0.00 [0.00, 0.00]
MAC	32/36	34/36	2.50 [-0.62, 6.25]

DBD's accuracy decrease is 6.25 [0.00, 18.75] percentage points and occurs on 1 image: 10 of its sixteen held-out predictions change from correct to incorrect. MAC's decrease is 2.50 [-0.62, 6.25] points; 3 images worsen and 1 improves. Neither interval excludes zero. The additional losses demonstrate sensitivity to the specified budget and search schedule on these examples; unchanged R-TPT accuracy and the four zero-accuracy rows do not establish convergence or robustness to a different search.

I Analysis identities and conditional uncertainty

For each family, the correctness masks are aligned by exact sample IDs and labels. The five K_i categories partition the image set; their weighted mean recovers MeanRA, their final category recovers WCRA, and the sum of exclusive attack contributions recovers the $K_i = 3$ category. For each attack, $\mathrm{RA}_a - \mathrm{Clean} = G_a - B_a$. These identities were checked for all 315 families, and all 1,260 attack accuracies agree with recorded runs.

Primary evaluations use one method seed, zero. The bootstrap uses seed 20260925 and 10,000 replicates. Within each of the fifteen fixed datasets it samples the original number of images with replacement, sharing indices across both methods, every available backbone, and all attacks. Each of the 45 tasks keeps weight $1/45$, with three backbones for each dataset. This equals averaging the three backbones within each dataset and then weighting the fifteen datasets equally. Percentile endpoints quantify conditional image-sampling uncertainty only. Neither dataset selection nor training/inference randomness is resampled. The six-dataset timing study is a distinct fixed cohort and is not expanded with server measurements.

J Sensitivity to benchmark composition

We recompute all seven methods' clean accuracy, MeanRA, WCRA, PGD-ℓ_2 encoded-cache accuracy, and clean-correct retention after omitting each of the fifteen datasets and each of the three backbones. Dataset omission leaves 42 tasks; backbone omission leaves 30. Each of these eighteen checks renormalizes equal weights over the retained backbone–dataset tasks. Table 18 reports their ranges. These are descriptive composition checks, not confidence intervals or generalization estimates for unseen datasets or new seeds. The PGD-ℓ_2 results retain the nominal pre-encoding radius interpretation described in the attack protocol.

Equal weighting of the fifteen datasets after averaging their three backbones is identical to weighting all 45 tasks by $1/45$. We verify this identity; it is not an additional sensitivity condition. All 315 method families contribute to the complete-cohort analysis.

All-method ordering. We compare all 21 method pairs against their complete 45-task ordering in each omission. Every pairwise ordering for MeanRA, WCRA, PGD-ℓ_2, and retention is preserved. Clean accuracy has 3 reversal events under dataset omission and 2 under backbone omission. MTA remains the clean-accuracy leader throughout. Comparisons use unrounded values with a 10^{-10} percentage-point numerical tie tolerance. Formed and resolved ties are tabulated separately. Some small WCRA and retention values coincide after rounding, so preserved numerical order does not establish practical importance or statistical significance (Table 19).

The MAC minus R-TPT MeanRA gap ranges from 1.60–2.59 percentage points after omitting a dataset and 0.03–3.84 after omitting a backbone. R-TPT's WCRA advantage ranges from 4.84–6.52 and 2.74–9.16, respectively. Without DataComp, the MeanRA gap is only 0.03 percentage points; preserving an ordering does not preserve the size of the difference.

Table 18: Composition sensitivity (%). LODO and LOBO are ranges over 15 leave-one-dataset-out and three leave-one-backbone-out aggregates. The recorded seed and input conditions stay fixed; ranges are sensitivity checks, not confidence intervals. PGD-ℓ_2 refers to the separately reported encoded-cache condition with nominal optimization radius one.

Metric	Method	45-task	LODO range	LOBO range
Clean	Zero-shot	67.81	65.85–70.32	65.93–70.33
	TPT	69.09	67.17–71.62	67.59–71.30
	R-TPT	68.51	66.60–71.22	67.40–70.61
	DBD	67.55	65.53–70.32	66.35–69.61
	MTA	**69.67**	67.80–72.15	68.06–72.07
	TTC	58.92	56.35–61.47	56.41–60.42
	MAC	66.90	64.88–69.38	65.04–69.06
MeanRA	Zero-shot	0.98	0.67–1.05	0.55–1.35
	TPT	15.92	15.36–16.47	15.42–16.46
	R-TPT	50.86	48.25–53.34	48.04–54.56
	DBD	**72.44**	70.73–74.65	71.36–74.59
	MTA	42.94	40.21–45.36	38.94–48.24
	TTC	20.77	19.39–21.67	20.05–21.97
	MAC	52.96	50.50–54.97	50.01–56.98
WCRA	Zero-shot	0.00	0.00–0.00	0.00–0.00
	TPT	0.00	0.00–0.00	0.00–0.00
	R-TPT	37.41	34.36–39.89	33.22–42.22
	DBD	**59.99**	57.58–62.67	58.51–62.74
	MTA	28.06	25.12–30.02	23.34–33.88
	TTC	0.49	0.39–0.52	0.33–0.62
	MAC	31.63	28.73–33.37	27.63–36.76
PGD-ℓ_2	Zero-shot	0.40	0.25–0.43	0.07–0.57
	TPT	0.53	0.31–0.57	0.18–0.75
	R-TPT	55.73	53.15–58.12	53.55–58.87
	DBD	**95.60**	95.30–97.30	95.05–96.27
	MTA	43.51	40.49–45.86	40.06–49.14
	TTC	5.92	4.54–6.34	4.82–7.08
	MAC	66.81	64.94–68.74	62.36–69.15
Retention	Zero-shot	0.00	0.00–0.00	0.00–0.00
	TPT	0.00	0.00–0.00	0.00–0.00
	R-TPT	50.82	48.46–53.95	46.21–55.80
	DBD	**82.52**	81.47–84.78	81.86–83.57
	MTA	36.54	34.01–39.05	31.21–43.22
	TTC	0.73	0.64–0.78	0.49–0.91
	MAC	43.72	41.44–45.70	38.36–49.61

Table 19: Strict pairwise reversals relative to the complete 45-task aggregate. Each cell gives reversal events / formed ties / resolved ties. LODO and LOBO test 315 and 63 comparisons per metric, respectively (all 21 method pairs). Comparisons use unrounded values with a 10^{-10} percentage-point numerical tie tolerance; displayed 0.00 values need not tie.

Metric	LODO	LOBO
Clean	3/0/0	2/0/0
MeanRA	0/0/0	0/0/0
WCRA	0/0/0	0/0/0
PGD-ℓ_2	0/0/0	0/0/0
Retention	0/0/0	0/0/0